HAL 9000, la Odisea del 2001

Davor Pavisic

Instituto de Investigación en Informática Aplicada
Universidad Católica Boliviana - Regional Cochabamba
e-mail: dp@ucbcba.edu.bo

El HAL 9000 de 2001

Hace más de 30 años, el escritor de ciencia ficción, Arthur C. Clarke (junto con Stanley Kubrick), se imaginó, entre otras cosas, que en este nuevo milenio existirían ordenadores inteligentes funcionando de manera análoga a los cerebros de los seres humanos. Tal es el caso del ordenador HAL 9000, protagonista de la novela *2001, una odisea en el espacio* [2], quien "razonaba" de forma muy similar (o igual) a la mente humana. El nombre de HAL, cuyas siglas curiosamente anteceden a las siglas de la conocida IBM, significa *Heuristically programmed ALgorithmic computer* y este significado lo relaciona directamente con el campo de la inteligencia artificial (IA). HAL era el sexto miembro de la tripulación a bordo de una nave con destino a Saturno[1] y sus responsabilidades incluían desde el soporte básico de vida en la "Discovery", navegación, comunicaciones, contingencias y emergencias de todo tipo, hasta compañía para el resto de la tripulación que se encontraba a bordo. Es interesante notar también que HAL era el único miembro despierto de la tripulación que conocía el objetivo real de la misión[2].

Durante los meses que dura el viaje, HAL demuestra poseer atributos casi humanos que le permiten establecer una relación profesional y de amistad con sus compañeros de viaje humanos. Tal era la similitud entre el cerebro de HAL y la mente humana que incluso la mente electrónica de HAL entra, y con justificados motivos, en un estado de paranoia como describe Clarke en la novela: *"Durante los últimos 150 millones de kilómetros,* (HAL) *había estado cavilando sobre el secreto que no podía compartir con Poole y Bowman. Había estado viviendo una mentira; y se aproximaba rápidamente el tiempo en que sus colegas sabrían que había contribuido a engañarles... el conflicto estaba destruyendo lentamente su integridad... el conflicto entre la verdad y su ocultación"* [2].

[1] Sin embargo en la película del mismo nombre, el destino final de la nave era Júpiter ya que los especialistas en efectos especiales de la época (años 70) no pudieron lograr imágenes de los anillos de Saturno lo suficientemente convincentes.

[2] Había tres miembros de la tripulación que se encontraban en estado de hibernación.

Clarke justifica el dilema de HAL indicando: *"todas las energías, poderes y habilidades de HAL habían estado dirigidas hacia un fin. El cumplimiento de su programa asignado era más que una obsesión; era la única razón de su existencia. Inconturbado por las codicias y pasiones de la vida orgánica, había perseguido aquella meta con absoluta simplicidad mental de propósitos... El error deliberado era impensable. Hasta el ocultamiento de la verdad lo colmaba de una sensación de imperfección, de falsedad... de lo que en un ser humano hubiese sido llamado culpa, iniquidad, o pecado. Pues como sus constructores, HAL había sido creado inocente; pero demasiado pronto había entrado una serpiente en su Edén electrónico"* [2].

Clarke describe la falla de HAL tal como le ocurriría a un ser humano: *"Había comenzado a cometer errores; sin embargo, como un neurótico que no podía observar sus propios síntomas, los había negado... éste era relativamente un problema menor; podía haberlo solucionado —como la mayoría de los hombres tratan sus neurosis— de no haberse encontrado con una crisis que desafiaba a su propia existencia. Había sido amenazado con la desconexión... con ello sería arrojado a un inimaginable estado de inconsciencia. Para HAL esto era equivalente a la Muerte. Pues el no había dormido nunca;... "* [2].

El razonamiento "humano" de HAL lo conduce a tomar decisiones drásticas: *"Así pues, se protegería con todas las armas de que disponía. Sin rencor eliminaría el origen de sus fustraciones... y después proseguiría la misión... sin trabas, solo."* [2]. La programación de HAL incluía contingencias como que HAL podría encontrarse en una situación semejante, es decir, "solo". En esta situación HAL estaba autorizado para tomar sus propias decisiones y continuar la misión según su propio "sentido común": *"... podía llegar el día en que HAL tomase el mando de la nave... y adoptaría las medidas que juzgara necesarias para la salvaguardia y la continuación de la misión... "* [2].

Clarke (como Kubrick) extrapoló muy alegremente a la Inteligencia Artificial del momento. Es del conocimiento de todos que un ordenador semejante no existe hasta ahora[3]. Pero, leyendo nuevamente la novela de Clarke encontramos que algunas de sus predicciones no fueron totalmente erradas: si bien HAL continúa siendo un personaje de ciencia ficción, veremos más adelante que la visión de Clarke en esa época, es decir, a finales de los años 60, estaba bien fundamentada para la predicción de semejante inteligencia artificial.

Clarke predijo acertadamente, entre otras cosas, el uso cotidiano de internet y su forma de navegación: *"... uno a uno conjuraría a los principales periódicos electrónicos del mundo; conocía de memoria las claves de los más importantes... ojearía rápidamente los encabezamientos.... permitiendole así leer con comodidad. Una vez acabado volvería a la página completa, seleccionando un nuevo tema para su detallado examen"* [2] y paradójicamente el deterioro de los contenidos de los medios de prensa: *"Cuanto más maravillosos eran los medios de comunicación, tanto más vulgares, chabacanos o deprimentes parecían ser sus contenidos. Accidentes, crímenes, desastres naturales y causados por la mano del hombre, amenazas de conflicto, sombríos editoriales... tal*

[3]Pese a que Clarke atribuyó el año (y el título de su novela) 2001 exclusivamente a Kubrick cuando este ya había fallecido [7].

parecía ser aún la principal importancia de las millones de palabras esparcidas por el éter" [2]. Clarke llegó incluso a imaginarse el funcionamiento de los hornos a microondas: *"Sus menús habían de ser simplemente abiertos e introducido su contenido en la reducida auto-cocina, que lanzaba un zumbido de atención cuando había efectuado su tarea"* [2]. Veamos ahora algunos detalles de su HAL 9000.

Los orígenes de HAL

En su novela, Clarke relata el supuesto "nacimiento" de HAL en enero de 1997: *"HAL era una obra maestra de la tercera promoción de computadores. Ello parecía ocurrir a intervalos de veinte años, y mucha gente pensaba ya que otra nueva creación era inminente"*. Clarke basa esta idea en el aún corto periodo de tiempo que llevaban existiendo los primeros ordenadores y la inteligencia artificial como rama de la informática: *"La primera* (promoción) *había acontecido en 1940 y pico, cuando la válvula de vacío, hacía tiempo anticuada, había hecho posible tan toscos cachivaches de alta velocidad como ENIAC y sus sucesores"*. Por esa misma época, en 1943, McCulloch y Pitts mostraron que una *red de neuronas artificiales* era capaz de llevar a cabo ciertas tareas computacionales relativamente complejas [8]. Un poco más tarde, en 1949, Donald Hebb propuso un modelo de aprendizaje para este tipo de redes neuronales [6]. Luego, en 1951, Dean Edmonds y Marvin Minsky construyeron una máquina electromecánica capaz de aprender, la cual incorporaba estas ideas. En 1954 Minsky, en base a los resultados obtenidos en redes neuronales, terminó su tesis doctoral *Theory of Neural-Analog Reinforcement Systems and its Application to the Brain Model Problem* [9]. En 1961, Frank Rosenblatt inventó el *Perceptrón* y desarrolló el *Teorema de Aprendizaje del Perceptrón* [13]. Casi al mismo tiempo, aparecía la *Adaline* (Adaptive Linear Network) de Widrow y Hoff [18].

Los cada vez más importantes resultados que se estaban logrando en el área de las redes neuronales artificiales, claramente impresionaron y entusiasmaron a Clarke, quien añade: *"Luego en los años sesenta habían sido perfecionados sólidos ingenios microelectrónicos..."* [2]. Clarke marca entonces la primera generación de los ordenadores desde su inicio hasta el estado del arte de la fecha (fines de los años 60).

Establecidas las bases históricas y con los resultados científicos obtenidos por los especialistas de la época en el área de la inteligencia artificial y las redes neuronales artificiales, Clarke plasma el pensamiento informático de la época en su novela: *"Con su advenimiento, resultaba claro que inteligencias artificiales cuando menos tan poderosas como la del Hombre, no necesitaban ser mayores que mesas de despacho... caso de que se supiera como construirlas. Posiblemente nadie lo sabría nunca; mas ello no importaba"* [2].

Clarke aprovecha entonces para extrapolar los avances de la informática y la inteligencia artificial mencionando a Minsky como uno de los pioneros en el área: *"En los años ochenta, Minsky y Good habían mostrado como podían ser generadas automáticamente redes nerviosas auto-replicadas, de acuerdo con cualquier arbitrario programa de enseñanza. Podían construirse cerebros artificiales mediante un proceso asombrosamente análogo al desarrollo de un cerebro humano. En cualquier caso, jamás se sabrían los*

detalles precisos; y hasta si lo fueran, serían millones de veces demasiado complejos para la comprensión humana" [2]. En esa época, Clarke no tenía la menor sospecha que Minsky, junto con su colega S. Pappert, en 1969 y tan sólo un par de años más tarde que la fecha de publicación de su novela "2001", se tornaría contra las redes neuronales desplazándolas a un periodo de "oscurantismo" que duraría más de diez años [11]. Por el contrario, Clarke se imaginó que las redes neuronales artificiales podrían llegar a imitar a los cerebros humanos en un tiempo relativamente corto, es decir, hasta la fecha de la creación ficticia de HAL en 1997: *"Sea como fuere, el resultado final fue una máquina-inteligencia que podía reproducir —algunos filósofos preferían aún emplear la palabra "remedar"— la mayoría de las actividades del cerebro humano, y con mucha mayor velocidad y seguridad"* [2].

Clarke extrapola también áreas de la inteligencia artificial como la visión artificial y la comprensión y generación del lenguaje hablado: *"La mayoría de las comunicaciones de HAL con sus camaradas se hacían mediante la palabra hablada. Poole y Bowman* (tripulantes) *podían hablar a HAL como si fuese un ser humano, y el replicaría en el... más puro inglés que había aprendido durante las fugaces semanas de su electrónica infancia*" [2]. Sin embargo, en cuanto a la capacidad de "pensar" de HAL, Clarke añade: *"Sobre si HAL podía realmente pensar, era una cuestión que no había sido establecida por el* (brillante) *matemático inglés Alan Turing en los años cuarenta..."*. Recordemos que Turing fue también uno de los pioneros en la inteligencia artificial. Además de ser autor del prototipo teórico de los ordenadores digitales, *La Máquina Universal de Turing*, gracias al trabajo de análisis de criptografía que realizó durante la II Guerra Mundial en los ordenadores Colossi[4] y su trabajo post-guerra en el ordenador ACE, Turing se convenció que los ordenadores, con el tiempo, adquirirían la capacidad de pensar. Fue Turing quien, con esta idea en mente, propuso la famosa *Prueba de Turing* para evaluar, de forma objetiva y clara, si un ordenador era capaz de pensar o no.

La prueba de Turing

Para evitar argumentos semánticos sobre las definiciones de palabras como *máquinas* y *pensar*, Turing propuso el *Juego de Imitación*. De forma simplificada, el juego se limita a tener un *interrogador* (humano) quien utiliza su *inteligencia natural* para distinguir entre las respuestas que le da un ser humano de las respuestas que le da un ordenador, mientras *interrogador* y *jugador* mantienen una conversación (a distancia) por medio de un teclado y una pantalla. Si el interrogador no puede distinguir a la *máquina* de la *persona*, entonces la máquina ha pasado la *Prueba de Turing* y, por ende, es inteligente. En términos de inteligencia, la prueba de Turing puede ser resumida como sigue: si la conversación que una persona mantiene con un ordenador (inteligencia artificial) es indistinguible de aquella que mantiene con un humano (inteligencia natural), entonces el ordenador está mostrando inteligencia [5].

Esta prueba posee una serie de ventajas transparentes sobre otras definiciones más complejas de inteligencia artificial: libera a la máquina de características antropomórficas y, además, es independiente de los detalles del experimento [5].

[4]Durante la guerra, Turing logró romper el complejo código de encriptación alemán *Enigma*.

El mismo Turing es llevado por el entusiasmo de la época cuando indica: *"Pienso que en aproximadamente cincuenta años será posible programar ordenadores, cuya capacidad de almacenamiento será de mas o menos* 10^9*, y hacerlos " jugar" el juego de imitación tan bien, que un interrogador promedio no tendrá más de un 70% de probabilidad de realizar la correcta identificación después de cinco minutos de conversación con el ordenador... pienso que al final de este siglo la opinion general de la gente se habrá alterado tanto que uno podrá hablar de máquinas que piensan sin esperar ser contradecido"* [5].

La proyección de Turing sobre la capacidad de almacenamiento es sorprendentemente acertada si interpretamos que su predicción se refiere al Gigabyte de capacidad de almacenamiento en disco que ahora se encuentra fácilmente disponible en el mercado. Es también frecuente entre los informáticos escuchar, a modo de broma, frases como: "...la computadora está pensando...". Más aún, una de las mentes más brillantes de nuestra época admitió, públicamente y sin pensarlo, que un ordenador había pasado la prueba de Turing. Esto ocurrió cuando Deep Blue, de IBM, logró una derrota devastadora frente a Garry Kasparov en la movida 36 de esta famosa partida de ajedrez hace un par de años. Kasparov simplemente no lo podía creer y quedó obsesionado con la idea que Deep Blue recibió "ayuda humana" durante la partida [7].

Es obvio que HAL fácilmente pasaría la prueba de Turing. El mismo lector de la novela llega a sentir emociones con respecto a HAL: de aprecio al principio y de adversión más adelante. Sin embargo, es claro que nuestra sociedad no está lista para recibir a "ordenadores inteligentes" similares a HAL. Evidentemente, la polémica que surge alrededor del tema no es simple y, quizá por eso, muchas veces eludida.

Las opiniones contrarias

Cuando los ordenadores aparecieron por primera vez, su objetivo final era simplemente realizar grandes cantidades de cómputos. Es por eso que son también llamadas "computadoras". Sin embargo, como lo afirma Minsky en [10], unos cuantos pioneros –especialmente Alan Turing– se imaginaron ordenadores que irían más allá de la aritmética y posiblemente lograrían imitar los procesos que se llevan a cabo dentro del cerebro humano. No es sorprendente que tales opiniones suscitaran la intensa oposición que se mantiene aún en nuestros días, tal como lo indica Minsky: *"la mayoría de las personas piensa que las computadoras no lo hacen"* [10]. Más aún, muchos expertos en informática afirman que las máquinas nunca lograrán pensar. Si es así, ¿cómo estas pueden ser tan inteligentes y, sin embargo, tan tontas? Veamos algunos de los puntos de vista típicos de los años 50 y las refutaciones hechas por Turing en la época.

La objeción teológica. *El pensamiento es una función del alma (inmortal) del hombre.* Turing refuta: *"me parece a mi que este argumento implica una seria restricción sobre la omnipotencia de Dios... ¿Acaso El no podría dar un alma a un elefante o a una máquina si así lo desease?"* [5].

La objeción matemática. Según el Teorema de Gödel: "...en cualquier sistema lógico lo suficientemente poderoso, se pueden formular proposiciones, las cuales no pueden ser probadas ni refutadas dentro del sistema...". *¿Cómo podría pues un ordenador, lógico por naturaleza, probar o refutar lógicamente sus propias proposiciones?* Turing acepta la validez de este argumento pero observa que tampoco existe prueba que el intelecto humano no sufra las mismas limitaciones [5].

El argumento de la conciencia. *No aceptaremos que una máquina es equivalente a un cerebro hasta que ésta logre escribir un soneto o componer un concierto debido a los pensamientos y emociones que haya sentido y no por simple casualidad* [5], *es decir no sólo escribir sino, también saber que ella (la máquina) lo ha escrito.* Sin embargo, este argumento nos lleva a la clásica posición según la cual "la única forma de saber cómo piensa una persona es ser esa persona". Pero antes de caer en argumentos circulares a los que lleva esta posición, Turing nota: *"es usual tener la cortesía de aceptar que todas las personas piensan"* [5].

La objeción de originalidad. *Las máquinas nunca hacen nada nuevo* o *nunca nos sorprenden.* Turing indica: *"Las máquinas me sorprenden constantemente... La mayoría de la gente que programa, con seguridad comparte esta experiencia"* [5].

La continuidad del sistema nervioso. *El sistema nervioso no es una máquina discreta.* Pequeñas variaciones en amplitud y fase de un impulso nervioso que llega a una neurona pueden ocasionar grandes diferencias en la amplitud y fase del impulso de salida de la neurona en cuestión. En consecuencia, es imposible reproducir el funcionamiento del sistema nervioso (continuo) con un sistema discreto. Turing no puede refutar de forma convincente este punto, sin embargo, los ordenadores de hoy en día pueden simular el comportamiento de dispositivos analógicos no-lineales con casi cualquier nivel de precisión (siempre y cuando se sepa de antemano la función de transferencia del dispositivo). Es interesante notar que estas características (la no-linealidad y la continuidad) son fundamentales para el funcionamiento de las redes neuronales artificiales, sistemas que ahora han mostrado ser capaces de resolver una extensa gama de problemas.

Muchos de estos argumentos son manejados aún en nuestros días. Actualmente, la mayoría de las personas acepta que los ordenadores pueden hacer muchas cosas, las cuales requieren que una persona piense para realizarlas. Entonces, ¿cómo puede ser que una máquina aparente pensar pero que en realidad no lo haga? Dejando de lado la pregunta de lo que es en realidad "pensar", la mayoría de nosotros respondería indicando que el ordenador está realizando una imitación superficial de la inteligencia humana. Es decir, ha sido programado para obedecer a ciertos comandos simples pero sin tener la más mínima idea de lo que está ocurriendo o haciendo, según el caso. Sin embargo, el problema es más profundo de lo que pensamos, ya que se tocan algunos conceptos sobre los cuales aun nosotros mismos no estamos de acuerdo en su definición.

[5] Aquí se implica al discutido tema del mono frente a una máquina de escribir quien, después de un número finito de intentos escribe una tragedia de Shakespeare.

La creatividad

La mayoría de la gente piensa que la creatividad requiere una especie de "don" mágico que simplemente no puede ser explicado. Por tanto, los ordenadores no pueden *crear* ya que, según piensa la gente, todo lo que hacen las máquinas puede ser explicado. Como dice Minsky: "Nosotros naturalmente admiramos a nuestros Einsteins y Beethovens y nos preguntamos si los ordenadores serán capaces algún día de crear tan grandiosas teorías o sinfonías" [10]. Minsky luego añade: "... pero debemos evitar caer en la trampa... No debemos mirar hacia los grandes logros de nuestra cultura antes de entender cómo la gente común hace cosas cotidianas y comunes. No pienso que haya una gran diferencia entre el *pensamiento cotidiano* y el *pensamiento altamente creativo*. No culpo a nadie por no poder hacer las cosas que hace la gente creativa y tampoco culpo a nadie por no poder explicar lo que es la creatividad. Sin embargo, no concuerdo con la idea que, porque no podamos explicar cómo funciona la creatividad ahora, nadie podrá imaginar nunca cómo funciona en realidad" [10].

Minsky, acertadamente señala nuestra ignorancia acerca de la creatividad cotidiana: "deberíamos estar molestos por nuestra ignorancia de cómo obtenemos nuestras ideas" [10], y argumenta contra la reacción común de la gente: si no lo podemos explicar, entonces lo atribuimos a algo divino y por tanto propio solamente al ser humano. Minsky, por el contrario, atribuye la creatividad a una consecuencia del aprendizaje.

Es, entonces, prácticamente imposible esperar que las máquinas logren hacer maravillas (como lo hace un Mozart) sin antes comprender cómo lograr que éstas hagan cosas ordinarias y simples como lo hacemos nosotros, la gente común.

En 2001, HAL demuestra poseer un alto grado de creatividad tanto por las conversaciones que mantiene con los miembros de la nave como por la formulación de un plan para evitar ser desconectado y poder continuar la misión que se la había asignado. ¿Cuán lejos están nuestros ordenadores de tener esta creatividad?

En 1950, Newell, Shaw y Simon desarrollaron el sistema *General Problem Solver* (GPS) [4]. Los objetivos de los autores era dos: (a) desarrollar un modelo operacional explícito de la forma en que los humanos resuelven problemas cotidianos y (b) implementar este modelo en un ordenador. La consideración principal en el diseño de este sistema era separar los métodos generales de resolución de problemas de los datos específicos del problema en sí [12].

Aquí comenzó una etapa de programas que podían resolver problemas utilizando un tipo de "razonamiento" que no podía ser anticipado por los propios programadores, ¿creatividad?. Estos sistemas no tomaban decisiones al azar, más bien utilizaban *consejos* sobre qué cosas o métodos podrían funcionar en una situación dada. Estos *consejos* eran dados por expertos en un área específica que, como dice su nombre, adquirieron habilidades por medio de la experiencia. Esta experiencia era pues transferida a las máquinas como lo hacemos nosotros en la vida cotidiana algunas veces: dando consejos.

Esta idea, junto con algunas otras, hizo surgir un método de programación que posteriormente evolucionó en los sistemas de producción y los sistemas expertos. Estos

programas automáticamente aplican reglas cuando es necesario y, contrariamente a la opinión de la gente, generan soluciones con mucha originalidad.

Por ejemplo, la versión precedente y similar en principios a GPS, el *Logic Theorist* logró probar 38 de los primeros 52 teoremas del capítulo 2 de *Principia Mathematica* de Whitehead y Rusell [17, 5]. Más aún, la prueba del Teorema 2.85 es, en realidad, más corta y más elegante que la prueba dada en el mismo *Principia Matemática.*

Sistemas similares comenzaron a surgir desde mediados de los años 60. Entre los más importantes podemos citar a DENDRAL que podía, al igual que un experto humano, identificar la estructura molecular de un compuesto a partir de su espectro de masa [5], MACSYMA, un sistema matemático capaz de realizar diferenciación e integración simbólicas, algebra matricial, soluciones de sistemas de ecuaciones, expansión de las series de Taylor, etc. [5], MYSIN, un sistema experto para el diagnóstico y recomendación de tratamiento para enfermedades infeciosas en la sangre [15], PROSPECTOR, un sistema experto para la toma de decisiones en la explotación de minerales [3].

Vemos pues que los ordenadores ya en los años 60 podían encontrar soluciones originales a una gran diversidad de problemas. Es más, en la actualidad, todos los sistemas de navegación espacial dependen exclusivamente de la capacidad de cálculo de los ordenadores. Sin embargo, tal vez, justamente para evitar la "creatividad de las máquinas", se utilizan algoritmos determinísticos muchas veces controlados desde los centros de control de tierra. Un programa que se encargue de la navegación de una nave ya no es ciencia ficción. Muchos de los aviones que utilizamos para desplazarnos de ciudad en ciudad están capacitados para efectuar esos viajes sin la necesidad de un piloto humano. No obstante, nuestros propios prejuicios nos evitan ceder la responsabilidad a máquinas que quizá podrían cometer algun error. Así pues, vemos que la tarea de navegación de HAL en 2001 puede ser realizada por cerebros electrónicos similares. Analicemos ahora si es posible trabajar con un ordenador utilizando como medio de comunicación el lenguaje natural, como lo hacía la tripulación de la "Discovery" con HAL.

El lenguaje natural

Pese a las dificulates que presenta la comprensión del lenguaje natural, se han logrado algunos éxitos significativos en el área. Uno de los más sorprendentes es la tesis doctoral de T. Winograd quien desarrolló SHRLDU en el MIT en 1972 [19]. Limitando el dominio de aplicación, con el uso de una representación procedural y utilizando una gramática libre de contexto, SHRLDU era capaz de integrar conocimiento sintáctico, semántico y de razonamiento.

SHRDLU simulaba un entorno en tres dimensiones en el que se encontraban dispuestas varias diferentes formas, tal como se puede apreciar en la Figura 1. SHRLDU podía manipular los objetos por medio de un brazo de robot. El operador humano daba las instrucciones al sistema SHRLDU en *lenguage natural* (inglés). Veamos una corta conversación entre el operador humano (normal), el robot (negrilla) y algunos comentarios de Winograd (cursiva). Esta conversación fue extractada y traducida de [5].

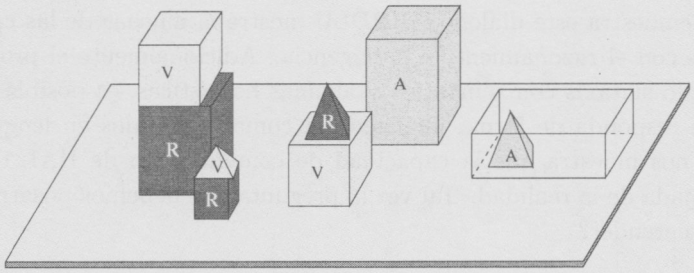

Figura 1: El mundo de bloques de SHRDLU. El operador humano puede comandar a un brazo de robot para mover los objetos verdes (V), rojos (R) y azules (A) usando lenguaje natural.

1. `levanta un bloque rojo grande`

   ok

   El sistema responde "OK" cuando realiza un comando exitosamente. Para poder levantar el bloque rojo, SHRLDU debía antes encontrar un espacio libre para el bloque verde y moverlo.

2. `toma la pirámide`

   no entiendo a qué pirámide te refieres

   Existen tres pirámides en el entorno y por tanto el programa no hace nada.

3. `encuentra un bloque que es más alto que aquel que tienes y colócalo en la caja.`

   El sistema debe darse cuenta a qué se refiere el operador por "aquel" como también utilizar la deducción para decidir cual es el bloque que "tiene".

   ok

4. `¿qué contiene la caja?`

   la pirámide azul y el bloque azul

5. `¿qué sostiene a la pirámide?`

   Nótese aquí el sistema supone que estamos hablando de la pirámide que se encuentra en la caja.

   la caja

6. `¿cuántos bloques no están en la caja?`

   cuatro de ellos

7. `¿es por lo menos uno de ellos más delgado que aquel que levantaste?`

   sí, el cubo rojo

8. `¿podrías, por favor, apilar ambos bloques rojos y un cubo verde o una pirámide?`

   ok

9. `¿tocaste alguna pirámide antes de colocar la verde sobre el cubo pequeño?`

   sí, la verde

10. `¿cuándo la levantaste?`

    cuando estaba apilando el cubo rojo, un bloque rojo grande y un cubo verde grande

Como lo demuestra este diálogo, SHRDLU mostraba muchas de las características que asociamos con el razonamiento e inteligencia. Adicionalmente el programa probó que combinando sintaxis con semántica y algunas heurísticas, es posible construir un programa que responda de forma inteligente a comandos dados en lenguaje natural. Este ejemplo nos muestra que la capacidad de comunicación de HAL tampoco está demasiado alejada de la realidad. Tal vez la pregunta que debemos posarnos es ¿puede un ordenador entender?

El entendimiento

En 1965 D. Bobrow concibió a STUDENT, un sistema capaz de resolver una serie de problemas de álgebra básica, cuyo enunciado se daba en lenguaje natural. Por ejemplo:

a) *La distancia entre Lima y La Paz es de 500 kilómetros. Si la velocidad promedio de un avión es de 900 Kilómetros por hora, encontrar el tiempo que toma viajar desde Lima a La Paz por avión.*

b) *El tío del padre de Juan tiene el doble de la edad del padre de Juan. Dentro de dos años, el padre de Juan tendrá tres veces la edad de Juan. La suma de sus edades es 92. Encontrar la edad de Juan.*

La mayoría de la gente concuerda en que este tipo de problemas es mucho más difícil de resolver que aquellos donde las ecuaciones matemáticas ya están dadas. Para resolver problemas con enunciados en lenguaje natural, uno debe encontrar las ecuaciones a resolver y, para lograr esto, uno debe entender qué es lo que las palabras y oraciones significan. ¿Acaso el sistema STUDENT entendía? En realidad STUDENT utilizaba una serie de "trucos". Por ejemplo, STUDENT estaba programado para "adivinar" que "**es**" generalmente significa "**es igual a**". Por otro lado, no intentaba siquiera saber lo que *el tío de Juan* significa —sólo se daba cuenta que esta frase se asemeja a *el padre de Juan.* Tampoco sabía que *edad* se refiere a tiempo pero la tomaba como algo que representa un número y que puede ser puesto en una ecuación. Con unos cuantos cientos de estos trucos, STUDENT, muchas veces, lograba encontrar las respuestas correctas.

Pero ¿podemos decir que STUDENT realmente entendía estos problemas? Minsky responde a esta pregunta de la siguiente manera: *"¿por qué molestarse? ¿por qué caer en la trampa de tratar de definir el significado de viejas palabras como "significar" y "comprender"?... La pregunta, más bien debería ser: ¿Acaso STUDENT trataba de eludir el "significado" utilizando algunos trucos?"* Vayamos más allá preguntándonos lo siguiente: es obvio que STUDENT sabe de aritmética en el sentido que puede encontrar una suma como "5 más 7 es 12". Pero, ¿puede STUDENT entender el significado del número? Por ejemplo, ¿qué "es" 5? o ¿qué es "más"? Con el afán de atribuir definiciones perfectas para palabras ordinarias, los filósofos B. Russell y A. North Whitehead, a principios del siglo pasado, propusieron una nueva forma para definir a los números: "cinco", decían ellos, es "la clase de todos los posibles conjuntos con cinco elementos". Sin embargo, mucha gente confunde con facilidad "clase" con "conjunto"

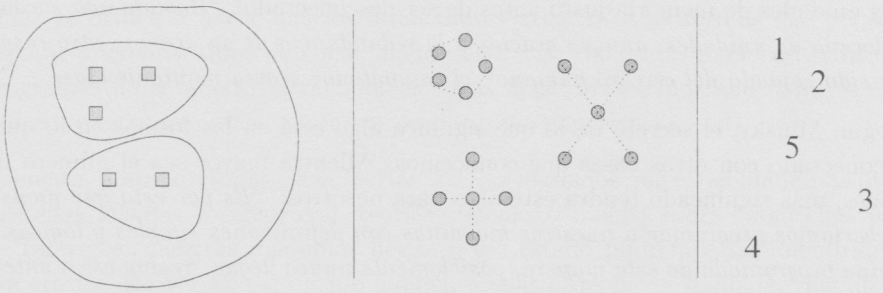

Figura 2: Diferentes formas de interpretar el Cinco.

y esta "pequeña" confusión lleva a inconsistencias que el mismo Rusell descubrió[6] [14]. El punto es que si bien muchos no pueden dar una definición precisa de número, esto no implica que desconocen el significado de, por ejemplo, Cinco. Según Minsky, "lo que significa algo para uno, depende, de cierta forma, en muchas otras definiciones que uno conoce... Por ejemplo, para el número Cinco podemos pensar en grupos de Dos y Tres, o Uno y Cuatro. También podemos pensar en algunas formas familiares: un pentágono, una cruz, una equis, etc. Todos forman Cincos." (ver Figura 2) [10].

¿Qué ocurriría si construimos máquinas que no estén basadas en definiciones rígidas? ¿No serían éstas llevadas a paradojas, inconsistencias y equivocaciones, como le ocurrió a HAL en 2001? Minsky, al respecto, indica: *"la mayoría de los conocimientos de la gente están llenos de contradicciones y, aún así, sobrevivimos... lo mejor que podemos hacer al respecto es ser razonamblemente cautelosos..."*. Minsky argumenta aquí que si deseamos que las máquinas lleguen a "razonar" como una persona lo hace, debemos darle los atributos necesarios: *"hagamos también a nuestras máquinas así de cuidadosas... y si existen algunas probabilidades de error, bueno, así es la vida"* [10].

Según Minsky, todas las definiciones que manejamos en el cerebro están relacionadas unas con otras en una gran *red de significados*. Debido a que cada persona tiene sus propias definiciones asociadas a otras, preguntarnos cuál es la correcta no tiene sentido. Cada definición tiene sus usos y sus formas de apoyar a otras definiciones. Ninguna tiene mucho poder por sí misma, pero juntas, hacen un sistema muy poderoso y versátil: *"... las redes en nuestras mentes son, probablemente, más complejas que cualquier otra estructura que la ciencia haya contemplado hasta ahora. Consecuentemente, la Inteligencia Artificial necesitará también, eventualmente, de algunas teorías extremadamente complejas. Pero ésa es también la vida"* [10]. Efectivamente, con un sistema como el propuesto por Minsky, cada palabra que nosotros utilizamos activa grandes redes con diferentes maneras de tratar y ver las cosas. Con redes de conocimiento masivamente conectadas no es posible atascarse, cuando un cierto significado falla, simplemente podemos utilizar algún otro hasta hallar el apropiado. Clarke también había pensado en esta propiedad en su HAL 9000, la cual se hace aparente cuando HAL es privado

[6]La paradoja de Rusell es frecuentemente ilustrada de la siguiente forma: *En un pueblo hay un barbero que afeita a todas aquellas personas que no se afeitan ellas mismas. La pardoja reside en la pregunta si el barbero se afeita o no* [1].

de sus unidades de memoria justo antes de ser desconectado: *"Habían sido sacadas ya una docena de unidades, aunque gracias a la redundancia de su diseño –otro rasgo que había sido copiado del cerebro humano– el computador seguía manteniéndose..."*.

Según Minsky, el secreto de lo que significa algo está en las formas en lo que éste está conectado con otras cosas que conocemos. Mientra mayor sea el número de conexiones, más significado tendrá este algo para nosotros: *"Es por esto que pienso que no deberíamos programar a nuestras máquinas con definiciones simples y lógicas. Una máquina programada de esta manera, posiblemente nunca llegue "realmente a entender" del mismo modo que una persona tampoco lo lograría. Cuando existen muchos significados en una red de conocimiento, es posible mover un poco las cosas en la mente y mirarlas con diferente perspectiva; cuando uno se estanca, es posible intentar otro punto de vista. Eso es lo que significa "pensar"... es por esto que prefiero redes de definiciones circulares. Cada una da significado al resto. No hay nada malo con gustar de varias diferentes melodías que contrastan unas con otras, —o nudos o tejidos— donde cada unidad ayuda a mantener a las otras juntas —o separadas—"*.

Minsky razona que, por supuesto, ninguna máquina logrará realmente comprender algo real o, incluso, saber lo que un número significa, si ésta es forzada a tratar con este algo en una única forma. Tampoco lo lograría un niño o un filósofo. Estas dudas no tienen nada que ver con los ordenadores, sino con nuestra tonta búsqueda de significados que "significan" solos, fuera de todo contexto. Nuestras preguntas sobre las máquinas que piensan deberían, en realidad, ser preguntas sobre nuestra propia mente.

El HAL 9000 en el 2001

Nuestras nociones de la mente humana son primitivas y, sin embargo, nos resistimos a admitir cuán poco sabemos sobre su forma de funcionamiento. Es posible que esto sea parte de nuestro sistema de autorepresión que nos evita pensar en problemas que aparentemente no tienen solución. Pero también existen razones más profundas: el deseo de creer en la unicidad y la inexplicabilidad del Ser.

Hay una ironía especial cuando la gente dice que las máquinas no tienen mente, ya que incluso ahora no hemos comenzado a entender cómo trabajan nuestras mentes. Sin embargo, ahora parece extraño que alguien pueda entender estas cosas sin tener un conocimiento más completo de lo que son las máquinas. Excepto, por supuesto, que se piense que las mentes no sean nada complicadas.

Para lograr máquinas inteligentes, necesitamos mejores teorías que determinen cómo "representar", dentro de los ordenadores, las redes de conocimiento y experiencia que se encuentran dentro del sentido común de la gente. Debemos desarrollar programas que "sepan" lo que significan los números, en lugar de ser capaz de simplemente sumar y restar. Debemos experimentar con todo tipo de conocimiento de sentido común.

Este es el foco de la investigación actual en Inteligencia Artificial. Es cierto que la mayor parte de las Ciencias de la Computación está dedicada a construir sistemas muy grandes y útiles pero nada "inteligentes". Sin embargo, una pequeña parte de

las mismas se dedica a hacer que los ordenadores utilicen otras formas de "pensar" y representar diferentes tipos de conocimiento en varias diferentes maneras, para que estos programas no se queden atrapados en ideas fijas. Más importante aún, se está intentando que estas máquinas aprendan de su propia experiencia.

Finalmente, juntando todas estas teorías, tal vez logremos que estas máquinas piensen sobre ellas mismas y construyan teorías (buenas o malas) sobre cómo ellas funcionan. Tal vez si nuestras máquinas llegan a ese estado podremos fácilmente decir que ha ocurrido. Tal vez en ese mismo momento ellas objeten ser llamadas máquinas. Minsky agrega al respecto: *"Aceptar esto será seguramente muy difícil, pero sólo con este sacrificio, las máquinas podrán liberarnos de nuestros falsas creencias"* [10].

¿Es posible, entonces, programar ordenadores que sean "conscientes" al igual que HAL? La gente generalmente espera que la respuesta a esta pregunta sea "no". Sin embargo, por lo menos en teoría, esto parece posible. Ya existen programas que poseen una inteligencia artifical limitada y que pueden "comprender" y ser "creativas" hasta cierto punto. Se han alcanzado ciertos logros en la comprensión del lenguaje natural y el hardware con el que se cuenta actualmente puede realizar tareas como el procesamiento de imágenes y reconocimiento del lenguaje hablado. Finalmente la robótica permite que los ordenadores puedan manipular su entorno. El problema radica en el poco conocimiento que tenemos para dar a los programas el suficiente sentido común que posiblemente sea necesario. Por ejemplo, es posible que sea demasiado arriesgado asignar a un ordenador una tarea importante a largo plazo sin darle antes una noción de sus propias habilidades (como por ejemplo la misión de HAL en 2001). No es deseable que éste empiece alguna tarea que no podrá terminar en un periodo aceptable de tiempo y, por tanto, sería importante que conozca sus propias limitaciones. En general, es posible que una máquina inteligente pueda entenderse a sí misma lo suficiente como para poderse cambiar y adaptar. Si esto ocurriese, durante un periodo de tiempo suficientemente largo, ¿por qué no podrían esas criaturas inteligentes evolucionar a un estado mental superior y, eventualmente, a un estado mental similar al nuestro?

Pasará todavía un buen tiempo antes de que aprendamos lo suficiente sobre el razonamiento del sentido común para lograr que las máquinas sean suficientemente inteligentes. Por el momento sabemos cómo crear sistemas expertos especializados y útiles pero todavía no sabemos cómo lograr que éstos sean capaces de mejorarse a sí mismos. Pero cuando logremos responder a estas preguntas, si lo logramos, tendremos que encarar una pregunta aún más extraña: Cuando sepamos cómo, entonces nos preguntaremos si debemos construir tales máquinas que, de algún modo, sean mejores que nosotros mismos. Por suerte dejamos esta opción a las futuras generaciones.

Minsky acertadamente indica al respecto: "De la misma manera que la Evolución ha cambiado el punto de vista del hombre hacia la Vida, la Inteligencia Artificial cambiará el punto de vista de la mente hacia la Mente. A medida que encontramos nuevas formas para que las máquinas se comporten con más sensibilidad, nosotros también aprenderemos más sobre nuestros procesos mentales". Es decir, encontraremos nuevas formas de pensar sobre nuestros "pensamientos" y "sentimientos" [10]. Nadie puede predecir a dónde nos llevarán estas ideas pero una cosa es segura por el momento: nuestra defi-

nición de las diferencias básicas entre las mentes de los hombres y las posibles mentes de las máquinas no están suficientemente claras. Quizá ya no sea posible establecer el límite donde la mente humana deja paso a la máquina. Quizá sea necesario que ellas nos den su opinión.

Referencias

[1] A. Bouvier y M. George, editores. *Dictionnaire des Mathématiques.* Presses Universitaires de France, 1979.

[2] A.C. Clarke. *2001, una Odisea Espacial.* J. Vergara, 1984. Primera Edición en 1968: 2001, a space odyssey - Polaris Productions.

[3] R.O. Duda y R. Reboh. *Artificial Intelligence Applications for Business*, Cap. AI and Decision Making: The Prospector Experience. Ablex Publishing Corp., 1984.

[4] G. Ernst y A. Newell. *GPS: A case study in generality and problem solving.* Academic Press, New York, NY, 1969.

[5] M. Firebaugh. *Artificial Intelligence, a Knowledge-Based Approach.* Boyd & Fraser Publishing Company, 1988.

[6] D. Hebb. *The organization of behavior.* Wiley, 1949.

[7] S. Levy. 2001: Why hal never happened. *Newsweek*, pp 52–56, Dic. 2000 - Feb. 2001. Special Edition.

[8] W.S. McCulloc y W. Pitts. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5:115–133, 1943.

[9] M. Minsky. *Theory of Neural-Analog Reinforcement Systems and its Application to the Brain Model Problem.* Tesis Doctoral, Princeton University, 1954.

[10] M. Minsky. Why people think computers can't. *AI Magazine*, 3(4), 1982.

[11] M. Minsky y S. Papert. *Perceptrons.* MIT Press, Cambridge, MA, 1969.

[12] A. Newell y H. Simon. *Human Problem Solving.* Prentice Hall, 1973.

[13] F. Rosenblatt. *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms.* Spartan Books, New York, NY, 1961.

[14] Rusell's paradox. http://plato.stanford.edu/entries/rusell-paradox, 1995. Stanford Encyclopedia of Philosophy.

[15] E.H. Shortlife. *MYCIN: Computer-based Medical Consultations.* Elsevier Press, 1976.

[16] G. Stix. 2001: Rating hal against reality. *Scientific American*, 284(1):26, 2001.

[17] A. Whitehead y B. Russell. *Principia Mathematica to *56.* Cambridge Univ. Press, 1962.

[18] B. Widrow y M.E. Hoff. Adaptive switching circuits. En *Western Electronic Show and Convention*, pp 96–104. Institute of Radio Engineers, 1960.

[19] T. Winograd. *Understanding Natural Language.* Academic Press, 1972.